The role of information in multi-agent learning

Eric Guerci, Mohammed Ali Rastegar

▶ **To cite this version:**

Eric Guerci, Mohammed Ali Rastegar. The role of information in multi-agent learning. 2009. halshs-00449536

HAL Id: halshs-00449536
https://shs.hal.science/halshs-00449536

Preprint submitted on 22 Jan 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GREQAM

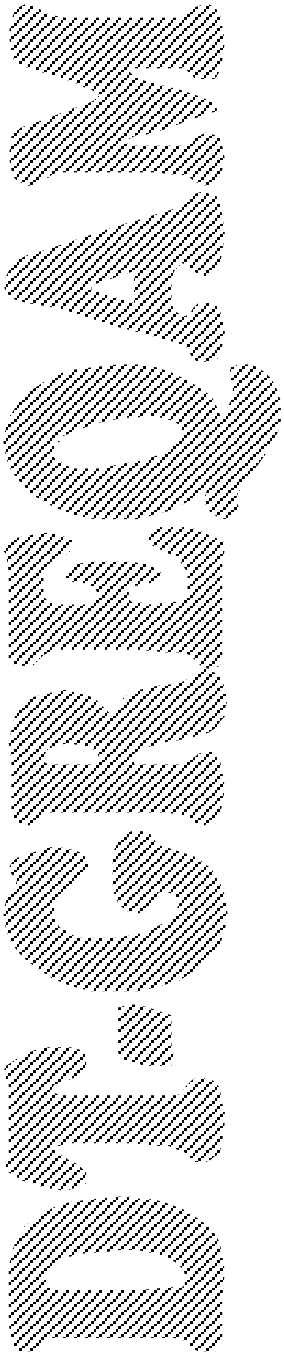
Groupement de Recherche en Economie
Quantitative d'Aix-Marseille - UMR-CNRS 6579
Ecole des Hautes Etudes en Sciences Sociales
Universités d'Aix-Marseille II et III

Document de Travail n°2009-47

THE ROLE OF INFORMATION IN MULTI-AGENT LEARNING

Eric GUERCI
Mohammad Ali RASTEGAR

November 2009



The role of information in multi-agent learning.

Eric Guerci[¶], Mohammad Ali Rastegar[§]

¶ *GREQAM, Université d'Aix-Marseille, 2 rue de la Charité, 13002 Marseille, France.*
§ *DIBE-CINEF, University of Genoa, Via Opera Pia 11a, 16146, Genoa, Italy.*

Abstract

This paper aims to contribute to the study of auction design within the domain of agent-based computational economics. In particular, we investigate the efficiency of different auction mechanisms in a bounded-rationality setting where heterogeneous artificial agents learn to compete for the supply of a homogeneous good. Two different auction mechanisms are compared: the uniform and the discriminatory pricing rules. Demand is considered constant and inelastic to price. Four learning algorithms representing different models of bounded rationality, are considered for modeling agents' learning capabilities. Results are analyzed according to two game-theoretic solution concepts, i.e., Nash equilibria and Pareto optima, and three performance metrics. Different computational experiments have been performed in different game settings, i.e., self-play and mixed-play competition with two, three and four market participants. This methodological approach permits to highlight properties which are invariant to the different market settings considered. The main economic result is that, irrespective of the learning model considered, the discriminatory pricing rule is a more efficient market mechanism than the uniform one in the two and three players games, whereas identical outcomes are obtained in four players competitions. Important insights are also given for the use of multi-agent learning as a framework for market design.

Key words: multi-agent learning; auction markets; design economics; agent-based computational economics

1. Introduction

Auctions are becoming a popular method for transacting business and the range of items sold by auctions has greatly increased in recent years due to e-commerce. In the last decade, auctions have also been considered to set up new markets, e.g., utilities and pollution permits markets. Accordingly, several theoretical studies about auction design have appeared in the economic literature during recent years, see Klemperer (2000); Milgrom (2004) for two recent monographs. In particular, economists have focused the attention to the design of efficient auction mechanisms for particular kinds of commodities, like electromagnetic spectrum (Cramton, 1998; Milgrom, 1998), carbon dioxide emission

rights (Cramton and Kerr, 2002), and electricity (von der Fehr and Harbord, 1993; Fabra, 2006).
This paper aims to contribute to the study of auction design from the perspective of Agent-based Computational Economics (ACE). In particular, we model trading activity by means of heterogeneous artificial agents characterized by different levels of learning capability and investigate the efficiency of two different double-auction mechanisms. The study of how agents can learn to compete or to coordinate in a market economy is a central issue in the ACE research agenda (Tesfatsion and Judd, 2006). In particular, the market design domain (Roth, 2002) is benefiting from ACE research (Marks, 2006). Indeed, the ACE approach can provide important insights about how agent behaviors and market settings influence one each other and determine the learning dynamics towards equilibria.
Many studies about auction design have recently appeared in the ACE literature, especially focusing on the design of electricity auction mechanisms (Nicolaisen et al., 2001; Bunn and Oliveira, 2001; Guerci et al., 2007; Yu et al., 2007; Sun and Tesfatsion, 2007; Guerci et al., in press). In particular, Nicolaisen et al. (2001) study market power and efficiency in a computational wholesale electricity market with discriminatory midpoint pricing, characterized by buyers and sellers which decide their orders according to a modified version of a well-known reinforcement learning algorithm (Erev and Roth, 1998). The authors show that, irrespective of the learning model's parameters, market efficiency depends on the market microstructure. Yu et al. (2007) analyze the day-ahead electricity market with locational marginal price; they compare a scenario where power suppliers are endowed with the Q-learning algorithm (Watkins and Dayan, 1992) with a scenario where suppliers have no learning capabilities and report their true marginal costs; the authors show that Q-learning suppliers are capable to make more profits in the long term. Guerci et al. (in press) analyze convergence properties of two reinforcement learning algorithms, i.e., the adaptive evolutionary algorithm proposed by Marimon and McGrattan (1995) and the Q-learning, in a duopoly and a tripoly economic scenario. The authors show that Q-learning agents are able to converge in the equilibrium of the infinitely repeated game.
The present paper stems from this previous strand of research and proposes an approach to auction design which aims to better encompass theoretical contributions and computational techniques originated within the theory of multi-agent learning (MAL). In order to appropriately to simulate market environments characterized by scarce information, we choose to endow agents with different degree of imperfect information about opponents' strategies. Both the Artificial Intelligence (AI) and economics communities have spent a lot of efforts in defining the problem of learning in a multi-agent context also with respect to the information available to the individual decision maker; accordingly, they have proposed taxonomies for defining appropriate classes of learning algorithms (Marimon and McGrattan, 1995; Chang and Kaelbling, 2001; Shoham et al., 2007). Coherently with the economic environment under study, we choose to employ a number of algorithms which belong to two standard classes of learning models, i.e., model-free and belief-based approaches. In particular, we use the Q-learning

(QL) (Watkins and Dayan, 1992) and the GIGA-WoLF (GW) (Bowling, 2005) developed within the AI domain, and the Marimon and McGrattan (MM) (Marimon and McGrattan, 1995) and the EWA-learning (EWA) (Camerer and Ho, 1999), which have been devised by economists. The above mentioned algorithms differ for what concerns the use of the agent's private information in the decision making process. MM, QL and GW belong to the model-free class, while EWA is an implicit belief-based model. Moreover, the fictitious play (FP) (Brown, 1951) algorithm, which is a pure belief-based model, has been also considered as a common opponent in order to select appropriate parameters for the previous four algorithms. Generally speaking, most of the theoretical and computational studies evaluate these algorithms in two-actions two-players games. In this respect, our contribution is the application and the evaluation of MAL in auction games with an increased number of players, which are characterized by a wider strategy space. Furthermore, we have devoted a great attention to properly evaluate both the convergence properties of the learning dynamic and the attainment of market efficiency. For this purpose, we draw the attention on convergence towards game-theoretic solution concepts, such as one-stage Nash equilibria and Pareto Optima, and performance metrics, such as profits and regrets. Indeed, we do not seek the best-performing algorithm, but, following the normative viewpoint of the game-theoretical approach, we investigate how different models of bounded rationality affect the attainment of market equilibria. This paper is organized as follows. Section 2 presents the economic environment under study. Section 3 introduces the issue of multi-agent learning and describes main features of the learning algorithms employed. The methodological approach to the design of computational experiments and the computational setting are described in Section 4. Section 5 presents and discusses results. Our concluding remarks are pointed out in Section 6.

2. Economic setting

Agents' strategy space

This paper studies an economic scenario characterized by the competition for the supply of a homogeneous good among a given number of producers. In the following, we use the terms agent, player, seller and producer interchangeably. We consider sellers deciding both price and quantity of their offer. Each ith agent submits a sell limit order which is characterized by a limit price p^i (ask price) and a corresponding quantity q^i. A finite two-dimensional strategy space $\mathcal{A}_i := \{(p^i, q^i) | 1 \leq q^i \leq \mathcal{Q}^i \quad \text{and } 0 \leq p^i \leq P^*\}$ has been considered for each agent. P^* is an upper bound for the price grid and $\mathcal{Q}^i$ is the maximum productive capacity for each agent. In order to increase results' intelligibility, the demand $\mathcal{Q}^d$ is assumed constant and inelastic to price.

Auction markets

Two different double-auction mechanisms have been considered: the uniform or system marginal price auction (UA) and the discriminatory or pay-as-bid

auction (DA). Their differ in the rule adopted to determine the clearing price between asks and bids.
In an uniform auction, the auctioneer builds the supply and demand curves and determines an unique market clearing price at curves' intersection. The supply curve is a discrete stepped curve defined by a price merit criterion, i.e.,

$$\mathcal{Q}(p) = \sum_{i|p^i \leq p} q^i \,.$$

Demand is constant and inelastic to price and thus can be represented by a vertical curve in the (q, p) plane. Ask orders are accepted if their prices are equal or lower than the clearing price. Accepted offers need to be rationed if their aggregate supply exceeds demand. Being focused on the decision-making process of sellers facing an inelastic demand, quantity rationing is significant only to the supply side of the market. In this respect, we adopt the standard approach of rationing the quantity only for offers with a price equal to the market clearing price. In particular, a quantity assignment problem arises when more than one seller, let's say n sellers, offer at the clearing price and $\sum_{i=1}^{n} q^i > \hat{\mathcal{Q}}^d$, where $\hat{\mathcal{Q}}^d$ is the residual demand given by $\hat{\mathcal{Q}}^d = \mathcal{Q}^d - \sum_{j=1}^{m} q^j$, being m the number of offers price below the clearing price. The rationing rule consists in subdividing the n sellers into two sets; A and B. The set A is composed by the n_A agents whose quantity offers q^i exceed the value $\hat{\mathcal{Q}}^d/n$, whereas the set B collects the remaining sellers. This equal rationing rule applies only for sellers in set A. The quantity traded by each seller in the set A, $\widehat{q}^i$ for $i \in A$, is then given by $\widehat{q}^i = (\hat{\mathcal{Q}}^d - \sum_{i \in B} q^i)/n_A$.

In the discriminatory auction, the matching procedure clears bids and asks progressively starting from the matching between the highest ask-price and the lowest bid-price. A transaction occurs at a price equal to the midpoint between ask and bid prices and at a quantity equal to the minimum between ask and bid quantities. The remaining ask or bid quantity is then matched with the second highest bid price or lowest ask price, respectively. This procedure is then iterated until there are ask prices equal or lower than bid prices; remaining offers are discarded. In the present study, because of the assumption of inelastic demand, i.e., undetermined bid-prices, the choice has been to set the transaction price at the accepted ask price. An indeterminacy arises in the matching procedure if two or more sellers offer at the same price and demand results lower than aggregate supply at that price. This situation is solved according to the equal rationing scheme adopted for the uniform auction mechanism.

Agents' profits (rewards)

We consider an economic scenario where production takes place only after sale, as in the electricity markets. Production costs thus depend only on the quantities $\widehat{q}^i$ which have been effectively traded. Stated constant and identical marginal costs c_m for each ith producer, profits (rewards, R) are then given as follows:

$$R^i = \left(p^i - c_m\right) \widehat{q}^i \qquad \text{(DA)}, \tag{1}$$

$$R^i = \left(P - c_m\right) \widehat{q}^i \qquad \text{(UA)}, \tag{2}$$

where, in the uniform auction case, P corresponds to the auction marginal price.

3. Multi Agent Learning theory

The first attempt to introduce and study learning in multi-agent systems was performed by game-theorists and dates back to the pioneer work of Brown (1951), which aimed to propose an algorithm for finding Nash equilibria. Since then, for many years on, a normative paradigm inspired this strand of research, where game-theorists were looking for bounded rationality models of players' behavior able to justify equilibrium concepts or to refine them (Fudenberg and Levine, 1999). In recent years, a descriptive approach has also been pursued by economists, motivated and justified by the widely-recognized experimental economics paradigm. They have been investigating behavioral justifications for an equilibrium theory, by explicitly estimating parametric models of learning on experimental data (Erev and Roth, 1998; Camerer, 2003). Curiously only in the last decade and quite independently, the topic of MAL has received increasing attention also from the Artificial Intelligence (AI) community. Specific application of robotic, distributed control problems and also entertainment/edutainment software motivated computer scientists to increase their research efforts in this direction. A computational and prescriptive goal is leading them in an attempt to define algorithms for "optimally" solving specific multi-agent systems tasks. In this respect, it is worth mentioning that, recently, a special issue about "Foundations of multi-agent learning" (Vohra and Wellman, 2007) has been published by the journal Artificial Intelligence. The special issue has been devoted to open a debate on the MAL agenda by bringing joint contributions of both the "machine learners" and economists' communities in order to highlight different viewpoints and experiences in the field. The starting point of the discussion is the paper by Shoham et al. (2007), where they attempt to pinpoint the goal of the research on MAL and the properties of the online learning problem. Five major lines of research are defined which encompass historical strand of research as well as new challenging ones. Normative, descriptive, computational and prescriptive categories are highlighted. In particular, according to the authoritative contributions by Fudenberg and Levine (2007) underlines, the theory of mechanism design can well benefit from development of computational techniques.

These important theoretical and computational considerations to the study of multi-agent systems can greatly inspire and justify the agent-based computational economics paradigm. They may provide ACE researchers with well-justified computational tools for investigating market economies. However, one might argue that for studying market rules, the learning approach could result an indirect method of computing equilibria and thus an inefficient solution. In recent years, there has been remarkable progress on developing direct techniques for equilibrium finding in normal form game (McKelvey et al., 2007; Sandholm

and A. Gilpin, 2005). These direct algorithms are more efficient in finding Nash equilibria and can certainly provide valuable tools for ACE researchers in market design. In this respect, it is worth remarking that in this paper we adopt both approaches. However, we believe that learning models are an approach which cannot be disregarded for addressing correctly the market design issue. Fudenberg and Levine (2007) stated that there is some reason to think that learning rules that are good rules from a prescriptive point of view may in fact be good from a descriptive point of view. Their viewpoint supports our approach to market design.

In this paper, we investigate different models of bounded rationality in the same market context in order to highlight market outcomes or invariance properties with respect of the learning models considered. In particular, we study five different types of learning agents and we test systematically them in two-players tournament and self-play competitions. Self-play competitions regard games where all players are endowed with the same learning algorithm. The two-players tournament encompasses both two-players self-play games and two-players mixed-play games, i.e., competitions occurring among different learning algorithms. The standard assumption is that each agent has no prior knowledge about the game structure or other player. Two classes of learning models have been considered. The first class refers to the so-called model-free approaches (Shoham et al., 2007) or reinforcement models of learning (Camerer and Ho, 1999), where agents learn a strategy that does well against the opponents without learning the opponents' strategies. The second complementary class regards model-based learning algorithms or belief-based learning models where players try to model opponents' strategies in order to play a best-response to them. The fictitious play algorithm, or even the Cournot best-response dynamics, can be seen as the ancestors of such class.
The different learning models studied in this paper have been selected in order to propose distinct models of learning representing both classes. This computational study includes the Marimon and McGrattan adaptive evolutionary algorithm (henceforth MM, see par. 3.4), the Q-learning algorithm (henceforth QL, see par. 3.3) and the GIGA-WoLF algorithm (henceforth GW, see par. 3.2) which belong to the first class and the EWA learning algorithm (henceforth EWA, see par. 3.1) and the classical fictitious play algorithm (henceforth FP, see par. 3.5) which belong to the second class. In particular, the MM and QL algorithms assume that individual agents are endowed with minimal information about the evolution of the game, i.e., they record only their own past history of plays and their associate instantaneous rewards. In this framework, agents do not know opponents' actions nor rewards, thus not having beliefs about alternative private paths of play, they reinforce only the last selected action. The MM and QL differ in the way the reinforcement process takes place. Indeed, the QL algorithm has been developed within the AI literature, like the GW, while the MM was conceived within the economics community. In particular, QL presents a temporal-difference mechanism (Sutton and Barto, 1998) which derives from considering the intertemporal discounted sum of expected rewards, originally conceived to solve model-free dynamic programming problems. The

GW learning model proposes a quite different learning approach, being based on a gradient-ascent technique which considers as target function the expected rewards. As far as concerns the second class, both EWA an fictitious play learning algorithm were developed by game-theorists. The FP is the oldest and is a pure example of belief-based learning model. The EWA is endowed with two important features. It is able to make hypothetical reasoning about alternatives plans of actions and rewards and is characterized by an implicit belief-based component.
A common probabilistic choice model has been adopted for the EWA and QL. The logit (exponential) quantal response function has been chosen in order to map the attractions/Q-values to a probability distribution function over actions. An important feature of the logit response function is that negative rewards can be taken into account. In particular, in this computational setting, we have assumed an increasing logit precision parameter $\lambda_t = \alpha t^{\beta}$ which increases with the simulation time step t. α and β are two positive and constant parameters, which have been appropriately tuned with respect to the different time-lengths of the computational experiments. Thereby, as the learning phase proceeds, the response functions become more responsive to propensities/attractions/Q-values differences, so agents are more and more likely to select better than worse choices. Accordingly, the very high value reached by the λ parameter at the final simulations' stages leads to a peaked probability distribution function on the strategy with the highest attraction/Q-value/propensity, i.e., the best strategy. This mechanism is intended to force the learning dynamics to converge to pure-strategies Nash equilibria. A similar probabilistic choice model has been implemented for the MM. New probabilities are determined exponentially weighting old probabilities with respect to updated propensities. The GW algorithm is the only algorithm which does not implement any probabilistic choice model because it computes strategies directly.
In the following, we introduce all algorithms, ordered according to the amount of information they deal with. However, we suggest to refer to the original paper for a detailed explanation and, if available, theoretical proof of convergence under restricted conditions such as self-play environment and characteristics of the game, e.g., zero-sum games, team games.

3.1. EWA-Learning (EWA)

The Experience-Weighted Attraction learning algorithm (Camerer and Ho, 1999) is a learning model which encompasses both reinforcement and belief-based learning models. Indeed, these are seen as two special cases of the more general EWA formulation. The key element of the EWA algorithm are attractions $A^i : \mathcal{A}^i \to \mathbb{R}$. Attractions are then monotonically related with the probability of choosing an action by considering an appropriate quantal response function which maps attractions A^i to strategies Π^i. The rule for updating attractions is the following:

$$A_t^i(a_j^i) = \frac{\phi N_{t-1} A_{t-1}^i(a_j^i) + [\delta + (1-\delta) I(a_j^i, a_t^i)] R^i(a_i^j, a_{-i}^j)}{N_t}, \tag{3}$$

where a_t^i is the last played action, $I(\cdot,\cdot)$ is the indicator function and N_t stands for number of "observations equivalents" of past experience which are updated according to:

$$N_t = \phi(1-\kappa)N_{t-1} + 1, \qquad t \leq 1, N_t \leq \frac{1}{1-\phi(1-\kappa)}. \tag{4}$$

For a detailed explanation of the three parameters δ, ϕ, κ, we suggest to refer to the original paper by Camerer and Ho (1999). However, we point out that the δ parameter may be considered as a tradeoff between the two classes of learning models considered. In particular, if $\delta = 0$ and $\kappa = 1$, the learning models coincides with a pure reinforcement learning model of learning, whereas if $\delta = 1$ and $\kappa = 0$, it coincides with a weighted fictitious play algorithm.
In this paper, we adopt the exponential (logit) rule for deriving probabilities from attractions:

$$\pi_t^i(a^i) = \frac{e^{\lambda_t A_t^i(a^i)}}{\sum_{a^i} e^{\lambda_t A_t^i(a^i)})}. \tag{5}$$

3.2. GIGA-WoLF (GW)

The GIGA-WoLF is an extension of the Infinitesimal Gradient Ascent learning algorithm (IGA) proposed by Singh et al. (2000). The idea of gradient-ascent techniques is to update mixed strategy in the direction of the current gradient of expected reward. The GW learning algorithm introduces two modifications to the simpler IGA version. The former (GIGA) refers to the generalization of the IGA algorithm which consists in considering two gradient-updated mixed strategies, $\pi^i(t)$ and $z^i(t)$ according to different steps size. This improvement allows to introduce a kind of "Win or Learn Fast" (WoLF) mechanism, that is, it learns faster if and only if its strategy π^i is losing to strategy z^i. The mathematical formulation follows:

$$\hat{\pi}_{t+1}^i = P(\pi_t^i + \eta_t r_t) \tag{6}$$

$$z_{t+1}^i = P(z_t^i + \eta_t \frac{r_t}{3}) \tag{7}$$

$$\delta_{t+1}^i = \min(1, \frac{\|z_{t+1}^i - z_t^i\|_2}{\|z_{t+1}^i - \hat{\pi}_t^i\|}) \tag{8}$$

$$\pi_{t+1}^i = \hat{\pi}_{t+1}^i + \delta_{t+1}^i(z_{t+1}^i - \hat{\pi}_t^i), \tag{9}$$

where $P(x)$ is an operator which projects the unconstrained vector x into the simplex of legal probability distributions:

$$P(x) = \arg \min_{\hat{x} \in PD(\mathcal{A}_i)} \|x - \hat{x}\|. \tag{10}$$

$\|\cdot\|$ is the standard $\mathcal{L}_2$ norm.
In this paper, we have adopted a variable learning rate $\eta_t = \alpha t^{-\beta}$ where α and β have been considered as the two positive and constant parameters of the learning model. Finally, One important and exclusive feature of this algorithm

is that it is a zero-regret algorithm that provably achieves convergence to a Nash equilibrium in self-play for games with two player and two actions per player. Please refers to the original paper of Bowling (2005) for further details.

3.3. Q-learning (QL)

QL algorithm is a popular algorithm for the single-agent framework, as Watkins and Dayan (1992) demonstrated the convergence to the optimal policy in the context of MDP. Nevertheless, the single-agent QL algorithm has already been adopted also for multi-agent learning problems (Littman, 2001; Hu and Wellman, 1998; Yu et al., 2007).
The ith agent takes an action a_t^i at time t and obtains a reward $R_t^i(a_t^i, a_t^{-i})$, depending also on the action played by the opponent a_t^{-i}. Then, it performs an update of the Q-function[1] $Q_t^i(a^i)$ according to the following recursive formula:

$$Q_{t+1}^i(a^i) = \begin{cases} (1-\alpha_t)Q_t^i(a^i) + \alpha_t[R_t^i + \gamma \max_{a'} Q_t^i(a')] & \\ & \text{if } a^i = a_t^i, \\ Q_t^i(a^i) & \text{otherwise;} \end{cases}$$

where γ is the discount factor and $\alpha_t = C_{t-1}^i(a^i)^{-\delta}$ is the variable learning rate. $C_t^i(a^i)$ is a counter function, i.e. a vector counting the number of times that an action $a^i \in \mathcal{A}^i$ has been played at time t since the beginning of the learning procedure. Then, the next action a_{t+1}^i, is randomly drawn from the probabilities determined by the exponential (logit) decision rule which states that:

$$\pi_t^i(a^i) = \frac{e^{\lambda_t Q_t^i(a^i)}}{\sum_{a^i} e^{\lambda_t Q_t^i(a^i))}}.$$

The convergence of $Q_t^i(a^i)$ to the optimal $Q^{i,*}(a^i)$ is guaranteed, only in the single-agent context, if α_t is appropriately decreased in time and each action is played an infinity of times. The Q-function thus corresponds to the expected total discounted reward for every action. Thus, the "optimal" policy is deterministic and results: $\pi^{i,*} = \arg\max_{a^i} Q^{i,*}(a^i)$. Thereby, Q-learning is capable of converging only to a pure strategy.

3.4. Marimon and McGrattan (MM)

Marimon and McGrattan (1995) propose an adaptive evolutionary learning algorithm where agents have minimal information about the evolution of the game. The mathematical formulation of the algorithm is the following: each seller i assigns a strength value $S_t^i(a^i)$ to every action a^i and updates it only for the played action a_t^i according to the realized profits, R_t^i, i.e.,

$$S_{t+1}^i(a^i) = \begin{cases} S_t^i(a^i) - \frac{1}{C_t^i(a^i)} \cdot [(S_t^i(a^i) - R_t^i(a^i))] & \\ & \text{if } a^i = a_t^i, \\ S_t^i(a^i) & \text{otherwise;} \end{cases}$$

[1]According to the infinitely repeated game framework, we model the multi-agent system with only one state thus determining a single state Q-function.

where $C_t^i(a^i)$ is the number of times that strategy a^i was played within the period of inertia (last revision of i 's mixed strategy $\pi_t^i(a^i)$) of the ith player, whose updating value is:

$$C_{t+1}^i(a^i) = \begin{cases} C_t^i(a^i) + 1 & \text{if } a^i = a_t^i, \\ C_t^i(a^i) & \text{otherwise.} \end{cases}$$

The inertia at auction round t is determined according to the parameter ρ, which establishes the probability of the ith player to update its strategy $\pi_{t+1}^i(a^i)$ at auction round $t+1$. The updating formula is:

$$\bar{\pi}_{t+1}^i(a^i) = \begin{cases} \pi_t^i(a^i) \cdot \frac{\exp(S_{t+1}^i(a^i))}{\sum_{a^i} \pi_t^i(a^i) \exp(S_{t+1}^i(a^i))} \\ \qquad\qquad \text{with probability } \rho, \\ \pi_t^i(a^i) \qquad\qquad \text{with probability } 1-\rho. \end{cases}$$

A peculiar feature of this algorithm is to have always a positive probability for every strategy. This mechanism is called experimentation and is described by:

$$\pi_{t+1}^i(a^i) = \begin{cases} \epsilon & \text{if } \bar{\pi}_{t+1}^i(a^i) \leq \epsilon, \\ \frac{\bar{\pi}_{t+1}^i(a^i)}{\sum_{a^i} \bar{\pi}_{t+1}^i(a^i)}(1-\bar{\epsilon}) & \text{otherwise;} \end{cases}$$

where $\bar{\epsilon} = \epsilon \cdot \text{card}[\pi_{t+1}^i(a^i) \leq \epsilon]$, $\epsilon \in (0,1)$. ϵ corresponds to the minimum probability value that can be assigned to any pure strategy.

3.5. Fictitious Play (FP)

Fictitious play is one of the first learning algorithm ever studied by game theorists. This non parametric learning rule is characterized by the fact that each agent presumes that its opponents are playing stationary mixed strategies and, recursively updating an estimate of their strategies' profile, it plays a best-response to them. The Cournot dynamics can be seen as a special case of this algorithm, when the estimation of opponents' strategies is done considering only the last play, in such a way disregarding all past information. Each ith agent iteratively updates a vector of beliefs (probabilities) $C_t^i(a^{-i})$ defined over the opponents' strategies space $\mathcal{A}^{-i}$ by cumulating their history of plays. In particular, every agent defines its beliefs normalizing the vector $c_t^i(a^{-i})$ counting the number of times the opponents played any own action till that time.

$$c_t^i(a^{-i}) = \begin{cases} c_{t-1}^i(a^{-i}) + 1 & \text{if } a^{-i} = \hat{a}_t^{-i}, \\ c_{t-1}^i(a^{-i}) & \text{otherwise;} \end{cases}$$

$$C_t^i(a^{-i}) = \frac{c_t^i(a^{-i})}{\sum_{a^{-i}} c_t^i(a^{-i})}. \tag{11}$$

Then, for next play, it chooses an action a^i which is a best response to those beliefs, i.e.,

$$\pi_{t+1}^i = BR_i(C_t^i) = \arg\max_{a^i} E[R_t^i(a^i, C_t^i)].$$

4. Computational Experiment Design

4.1. Methodological issues

Results provided in this paper are obtained by means of computational experiments. A common criticism toward the computational approach with respect to the analytical one is that it provides results with lack of generality. Indeed, simulation experiments often are the only feasible approach to deal with complex systems. Multi-agent systems are a typical example of complex systems where complexity is further increased by the adaptive behavior of learning agents. Multi-agent learning is characterized by the features of non-stationarity and heterogeneity, which make the study of convergence behavior very difficult to tackle analytically and leave the computational approach as a promising alternative.

This paper studies a multi-agent economic system by exploring how its properties vary with respect to different settings and agents' behavioral models, with the aim to increase the generality of our computational insights. This approach requires to perform a high number of computational experiments. In particular, we want to study the outcomes of two common auction mechanisms, with respect to different degrees of learning capabilities and information available to agents. We consider a limited number of agents, i.e., two, three or four, on the supply-side, each endowed with individual models of learning. We address the heterogeneity of agents behavior by considering the interaction among four different learning behaviors in a two-players tournament (2T) and three and four self-play competitions (3S and 4S). This is a distinctive methodological approach with respect to studies in the ACE literature, where only one algorithm is employed, see e.g. Yu et al. (2007) and with respect to the AI domain where different learning algorithms are tested with the major purpose of measuring their relative performance, see e.g. (Lipson and Leyton-Brown, 2005). Indeed, our methodology is to consider different learning behaviors, both in the self-play and in the mixed-play settings, in order to verify the existence of properties of the economic system under study which are invariant with respect to the behavior of agents. It is worth noting that we consider an environment where a fixed and limited number of heterogeneous agents play repeatedly each other, instead of addressing a large-population model, such as the anonymous random matching model (Fudenberg and Levine, 1999), which has been also adopted in the learning in games literature. The rationale for this choice is two-fold. First, our setting refers to relevant market scenarios (e.g. the electricity market) which adopt double-auction mechanisms and are characterized by the repeated interactions of the same market actors. Second, our analysis aims to consider a solution concept, that is the equilibrium in repeated games, which is not applicable in some environments with a large population of agents (Fudenberg and Levine, 2007).

The four learning algorithms considered in this paper are parameters dependent; the appropriate selection of parameters is a critical point and may be related to the particular setting where the learning algorithm is employed. Indeed, in

our framework the algorithms face many different settings which vary with respect to the game type, the kind of the algorithm used by the opponents and the number of players. The heterogeneity of our market environment poses a serious problem on the definition of a suitable parameters' selection procedure. Our approach has been to define a common environment for the selection procedure and aims to provide a common past experience to all learning agents. The selection procedure has been performed in two-player games with a common learning algorithm as opponent, both in the UA and the DA auction settings. The common opponent is the classical fictitious play algorithm, which has the advantage of being parameter-free. Parameters have been selected among a grid of possible values according to the criterium of the best convergence properties towards Nash equilibria in pure strategies. Parameters values, determined in the two-players competition with the FP, have been also employed in the 3S and 4S games. This choice is in accordance with the stated methodological approach to endow each agent with a common experience to be exploited in different market settings both with respect of the number of opponents and their learning capabilities. This approach can shed lights on robustness of the learning model.

Computational setting

We performed different computational experiments consisting in a two-agents tournament (2T) and self-play competitions with three (3S) and four (4S) agents. Self-play competitions are characterized by agents endowed with the same learning algorithm, whereas the tournament considers both mixed- and self-play two-players games.

Each agent is characterized by a two-dimensional strategy space (see Section 2) where the price grid upper bound P^* is set to 3 and maximum productive capacity $\mathcal{Q}^i$ is equal to 4. Both prices and quantities are measured in discrete units, thus determining a strategy space composed by 16 pure strategies a^i for each seller. Demand is equal to $\mathcal{Q}^d = 4$ in all computational experiments. Marginal costs are equal to $c_m = 1$ for each producer. The units are arbitrary.

As far as concerns two-players games, the joint strategy space consists in a set of 256 vectors of strategies. The joint strategy space results in a set of 4,096 and 65,536 vectors of strategies for the three- and four-players games, respectively.

Five different metrics have been employed to investigate the long-run behavior of learning experiments. In particular, we have considered three performance measures, i.e., profits (rewards), average regrets and average incentives to deviate, and two game-theoretic solution concepts, i.e., one-stage game Nash equilibria in pure strategies and one-stage game Pareto optima. See Appendix A for details about each proposed metric.

Profits are worth to be considered being included in the objective function of each learning algorithm; profits have also a high economic meaning with respect to market efficiency considerations. Average regrets and average incentives to deviate are taken into account because they give meaningful information about algorithm behavior out of equilibrium.

Convergence behavior has been studied considering pure-strategies. Pareto optima are considered for two reasons. First, it is an equilibrium solution concept

Agents in game	Auction	Nash	Pareto	Nash-Pareto
2	DA	14 (*5.5*)	13 (*5.1*)	9
	UA	17 (*6.6*)	67 (*26.2*)	16
3	DA	90 (*2.20*)	63 (*1.54*)	0
	UA	471 (*11.50*)	783 (*19.12*)	240
4	DA	2118 (*3.23*)	1152 (*1.76*)	0
	UA	8077 (*12.33*)	5584 (*8.52*)	1472

Table 1: Occurrences and relative percentages $(\cdot)$ of one-stage game Nash equilibria in pure strategies, one-stage game Pareto optima and joint Nash equilibria and Pareto optima.

Auction	EWA			GW		QL		MM	
	ϕ	δ	κ	α	β	δ	γ	ρ	ϵ
DA	0.9	0.4	0.9	0.96	-3	0.6	0.98	0.07	0
UA	0.9	0.4	0.9	0.54	-3	0.65	0.95	0.13	0

Table 2: Parameters values for the four learning algorithms considered in each auction mechanism. Parameters have been assumed constant in the 2T, 3S and 4S computational experiments.

in infinitely repeated games; second, Pareto-dominance among the set of Nash equilibria is an important refinement (Gordon, 2007).
Table 1 shows the number of one-stage game Nash equilibria in pure strategies and one-stage game Pareto optima for the two auctions considered and for different numbers of game players. The Table shows that every game has multiple Nash equilibria in pure strategies. Furthermore, the Table highlights an important difference between the two auction games. In the DA mechanism, an increase of the number of players to 3 and 4 leads to the absence of joint Nash equilibria and Pareto optima. Conversely, in the UA mechanism, joints solutions still exist in the case of 3 and 4 players. Table 2 reports the parameters values for each algorithm considered. The listed parameters have been selected according to the procedure described previously in this section. Values have been assumed constant in the 2T, 3S and 4S computational experiments. Initial probability distribution among actions has been assumed uniform for all algorithms. In the EWA, QL, MM algorithms, this assumption has been realized by setting the initial value of attractions/Q-values/propensities equal to zero.

5. Computational Results

Computational experiments results are reported in Tables 3-7. Each Table refers to a particular metric and reports results for both the UA and DA mechanisms; values within square brackets $[\cdot]$ refers to UA while normal brackets $(\cdot)$

regards DA. The first five columns of results regards 2T experiments. In particular, the first four columns report performance values of row algorithms against column opponent; the fifth column reports the average of 2T performances computed considering row values of previous four columns for both two-players self- and mixed-play games. FP column presents results obtained in the parameters' selection procedure, for which the convergence towards Nash equilibria was used as optimality criterion. Furthermore, the last three columns show results for the row algorithm in two-, three- and four-players self-play games. As far as concerns Tables 3, 6 and 7, the reported metrics values are averaged over the two-, three- and four-agents. Finally, results obtained by the FP algorithm are reported at the bottom of each table for the sake of completeness.
For every experiments, performance values and game theoretic solutions' frequencies have been estimated in the final part of the simulation as ensemble averages over sets of 100 independent simulation run. Values are reported with two-digit precision because of standard errors below 0.01. Each set of 100 simulation runs refers to a particular auction design. Simulation runs consist of 3,000 rounds for two-players simulations; the number of rounds is increased to 50,000 and 700,000 rounds for three-players and four-players games, respectively. The rationale for this choice is to increase proportionally the number of rounds according to the dimension of the joint strategies space.

Our major economic result is that the DA is generally a more efficient auction mechanism than the UA, irrespective of the learning paradigm considered. This finding emerges from Table 3 which shows that profits over two- and three-players games exhibit higher values for the UA mechanism. In this respect, it is worth noting that the demand of the representative buyer is always satisfied. This property is independent of the learning algorithm considered and thus on the degree of information made available to players. The difference between the two auction mechanisms diminishes with the increase of competitors' number; in particular the DA and UA present identical profits for all algorithms in the four players' case. This interesting economic finding can be better understood according to the following equilibrium analysis, which considers game-theoretic solution concepts such as one-stage Nash equilibria in pure strategies and one-stage Pareto optima.
Computational results show that the best convergence properties towards Nash equilibria are obtained by the EWA algorithm. This finding is clearly evident in the DA case on both 2T averages and self-play games. Besides, very good convergent properties are also present in the UA mechanism, where the Nash frequency value of EWA is below the GW only in the 2T average. Generally speaking, it is worth noting that the frequency of Nash equilibria is affected by the increase of the number of players; in particular, 3S and 4S games are characterized by more competitive market outcomes, with lower difference among algorithms and higher frequency values for Nash equilibria.
Worst convergence properties towards Nash equilibria are exhibited in the DA mechanism by the GW algorithm, except for the 3S and 4S cases. An analysis of average incentives to deviate provides a way to interpret this outcome. In

particular, a near zero incentive to deviate for the GW and a non-zero value for its opponents indicate that the GW is able to play a best response to opponent strategy, but not viceversa. It is worth remembering that the EWA, QL, and MM algorithms are characterized by a model of probability updating which leads in the long-run to the selection of a stationary action corresponding to the maximum expected payoff. Besides, an accurate analysis of computational results shows that the GW adapts to the convergent behavior of its opponent by estimating a final mixed strategy which establishes positive probabilities only to all best-response actions with respect to the stationary opponent's action. Therefore, even if GW always plays a best-response, the final stationary strategy of the opponent is the best-response only to one of pure-strategies selected by the GW agent. This fact leads to the low frequency values in two-players mixed-play games of Nash in pure-strategies. Similarly, relatively low frequency values occur also in the GW two-players self-play games, but in this case, the outcome is due to GW convergent behavior in self-play to select Nash equilibria in mixed-strategies which are not taken into account in Table 1. However, the GW is characterized by near-zero values of the incentive to deviate in all market settings and results to be the best performing algorithm also with respect to the average regret metrics. This latter feature is in accordance with the theoretical proof of learning with zero regret for the GW in the two-players self-play case, and extends its validity from a computational point of view for both two-players mixed-play games, and three- and four-players self-play games.

As far as concerns the Pareto solution concept, an interesting issue regards the ability of the learning algorithms to converge to refined Nash equilibria according to Pareto dominance. Table 1 shows that all market settings, except for the three- and four-players DA games, are characterized by a number of Nash equilibria which are also Pareto optima. Generally speaking, a learning algorithm with a 100 percent convergence to Nash equilibria is able to refine Nash equilibria according to Pareto-dominance if it contemporary exhibits Pareto optima frequency value close to one. Indeed, Table 5 reports frequency values significatively lower than one in different market settings. As far as concerns the two-players DA game, low frequency values of Pareto optima are generally observed in spite of a frequency of Nash equilibria close to one. According to Table 1, which reports 9 joint Nash and Pareto solutions out of a total of 14 Nash equilibria, a random selection of Pareto optima among Nash equilibria would give a frequency around 60 percent for Pareto optima. EWA, QL and MM exhibit values not far from this percentage value, while GW shows far lower values, highlighting an unexpected tendency to select Nash equilibria which are not Pareto. Conversely, in the two-players UA game, the algorithms' refinement capabilities can not be settled, because in this case almost all Nash equilibria are also Pareto optima, as reported in Table 1.
A striking no-refinement outcome results considering the three- and four-players self-play UA games, where a common behavior can be observed among all algorithms, i.e., the selection of Nash which are not Pareto. In particular, GW algorithm shows a pronounced tendency of no-refinement ability. However, it is

worth noting the outcome of the EWA learning algorithm in the three-players UA self-play game; in that case, Table 5 presents a frequency value of Pareto optima higher than the random selection value around 50%, see Table 1, and, accordingly, Table 3 reports profits far higher than other algorithms. Partial refinement ability can thus be attributed to the EWA learning algorithm only in the three players case.
These findings point out an important consideration: irrespective of the model-free or belief-based learning models considered, similar coordination failures arise. The learning agents are unable to coordinate in the log-run their strategies in order to achieve a strictly Pareto-dominant equilibria. The rationale for this finding might be related to the specific parameters' selection procedure adopted. However, even the FP algorithm, which does not require any parameters' selection procedure, shows identical convergent results. This result highlights a relevant invariant properties of the two classes of bounded-rationality models adopted, see 3. The higher degree of information available to players which characterize the belief-based algorithms seems unable to make Pareto-dominant outcomes feasible.

According to Table 1, no joint Nash equilibria and Pareto optima exist in the 3S and 4S DA games. Thereby, the refinement according to Pareto-dominance is not feasible in these market settings. However, it may be interesting to look for solutions which are are equilibria in the repeated game framework. In this respect, one-stage Pareto optima can be considered. Indeed, Table 5 shows that no algorithm is capable to learn this "tacit collusive" outcome. In particular, even the QL algorithm, which might have been the best candidate due to its intertemporal optimization behavior in the single-agent framework, is unable to learn tacit collusion. Therefore, competition does prevail for all algorithms in all market settings, also with a discount factor (γ) equal to 0.98 for the QL algorithm.

A final remark concerns the robustness of the learning models with respect to different game settings. Parameters have been determined in a two-players game with respect to a common opponent, for every algorithm and separately for each auction mechanism. However, a number of invariant properties characterizing each algorithm can be observed. In particular, the EWA algorithm exhibits the best convergent properties irrespective to the number of players and the auction mechanisms considered; indeed, it is the only algorithm characterized by the same set of parameters selected for both auction mechanisms, see Table 2. This property holds in particular in self-play games and also in mixed-play games, where performances, even if lower than the self-play setting, are greater than the ones reported by other algorithms in the same setting. Indeed, a common finding is that the heterogeneity of learning models in mixed-play competitions affects significatively coordination for convergence to an equilibrium. As far as concerns GW algorithm, a robust and invariant property is the general very good performance of both average regret and incentives to deviate, see Tables 6 and 7. An interesting remark is that this good result is obtained

by an algorithm which does not form any beliefs about opponents' strategies. Conversely, with respect to these latter metrics, QL and MM algorithms exhibit the general worst performance. A possible explanation relies on the simpler structure of their learning model. However, the convergent behavior of such learning rules improves when the number of players increases and in particular coincides with the best performing EWA and GW in the four players games.

6. Conclusions

The computational study of multi-agent systems offers a new interesting framework to market design. In a bounded rationality and incomplete information perspective, market outcomes may depend on behavioral models of economic agents. Multi-agent learning theory aims to provide a conceptual framework for modeling and analyzing bounded-rationality behavioral models in a heterogenous and interacting agents setting. The complexity of the problem under study usually induces the researcher to look at the computational approach as promising. Indeed, around ten years ago, the distinguished economists Fudenberg and Levine stated that in the learning in games domain "Laboratory experiments may not be perfect tests of theories but they are much better than no tests at all". We think that nowadays this sentence could be rephrased mentioning computational experiments besides laboratory ones.
This paper aims to contribute to establish a computational approach to market design consisting in studying the properties of particular market settings according to different behavioral assumptions about market participants. In particular, four different learning paradigms about agents' behavior have been used for studying the efficiency outcomes of two auction mechanisms, which are commonly employed, e.g. in the design of new deregulated electricity markets. Our methodological approach consists in performing several computational experiments characterized by mixed- and self-play competitions with an increasing number of game participants. Particular attention has been also devoted to parameters selection of learning algorithms. This approach is aimed to strengthen the reliability of computational results by highlighting the invariant properties of the market setting under study with respect to an increasing number of market participants, each endowed with different learning capabilities.
Our major economic result is that, irrespective of the learning algorithm considered, the DA is a more efficient auction mechanism than UA in the two- and three-players game setting. Indeed, another relevant property, which is invariant with respect of the bounded-rationality model considered, comes out from our computational experiments. The difference between the two auctions diminishes when the number of players increases. In particular, in the four-players game setting, the long-run convergence behavior of the learning dynamics is identical among the four algorithms considered. This results is particularly relevant for the UA mechanism where the selected equilibrium is not refined according to Pareto dominance. In this case, a coordination failure arises for both model-free and belief-based learning models. It is worth noting that this finding occurs also for a pure belief-based parameter-free model such as the fictitious play. These

results point out relevant indications for the design of new markets with respect to the information available to market participants.
In this paper, interesting insights are also obtained concerning the multi-agent learning domain and its usage as a framework for market design. The different degrees of information which characterize the four learning algorithms considered has an important influence on their performance and on the convergence to market equilibrium. In particular, the EWA learning algorithm emerges as the best algorithm to study market equilibrium outcomes in pure strategy, because it exhibits the best convergence properties according to Nash equilibria and Pareto optima. The EWA algorithm shows robust results with respect to different game settings and presents a good coordination ability with the opponent both in self-play and mixed-play competitions. Indeed, among the algorithms considered, the EWA is the most sophisticated, being the only one to implicitly form beliefs about the strategies of the opponents. The GW algorithm also exhibits very goods results in terms of average regrets and incentives to deviate, in accordance to its formulation as a no-regret learning algorithm. However, its convergence properties are not as satisfying as for the EWA both in self- and mixed-play. Indeed, the GW has been conceived to converge to Nash equilibria in mixed strategy, which are not considered here; besides, the very low values of average regrets and incentives to deviate are an indication of its ability to always play a best response with respect to the long-run stationary strategy of the opponent. GW exhibit the lowest refinement ability with respect to Pareto dominance; this latter fact lead to the worst average profits among the algorithms considered. Finally, the QL and MM algorithms are the least robust learning models with respect to the different game settings considered. In particular, the QL algorithm, even if endowed with intertemporal optimization capabilities, seems unable to learn Pareto solutions which are equilibria in the infinitely repeated game framework.

Three major interesting future lines of research are opened by this study. The first regards the adoption of more sophisticated learning algorithms, e.g. models which explicitly form beliefs about opponents behavior. The second concerns with the possibility to take into account convergence to equilibria in mixed strategies. In this respect, the development of more powerful equilibrium-solver techniques would help the interpretation of results. The last line of research refers to the study of more realistic market scenarios, e.g. characterized by a higher number of learning agents in both market sides. As a final remark, an interesting methodological approach would regard the exploitation of synergies between computational and laboratory experiments in order to support the validity of results in both frameworks.

	EWA	*GW*	*QL*	*MM*	*avg*	*FP*	*2 pl.*	*3 pl.*	*4 pl.*
EWA	[4.00]	[4.92]	[1.99]	[5.43]	[4.08]	[4.88]	**[4.00]**	**[2.25]**	**[1.00]**
	(3.21)	(2.32)	**(3.22)**	**(3.58)**	**(3.08)**	(3.24)	(3.21)	(1.33)	**(1.00)**
GW	[3.08]	[3.99]	[2.00]	[3.52]	[3.15]	[5.36]	[3.99]	[1.33]	**[1.00]**
	(2.33)	**(2.62)**	(1.99)	(2.14)	(2.27)	(2.04)	(2.62)	(1.33)	**(1.00)**
QL	**[5.97]**	**[5.60]**	**[3.00]**	**[5.58]**	**[5.04]**	**[6.00]**	[3.00]	[1.47]	**[1.00]**
	(3.20)	(2.09)	(2.90)	(3.51)	(2.92)	**(3.56)**	(2.90)	(1.33)	**(1.00)**
MM	[2.49]	[4.48]	[2.02]	[3.94]	[3.23]	[5.46]	[3.94]	[1.67]	**[1.00]**
	(3.54)	(2.14)	(3.17)	(3.28)	(3.03)	(2.04)	**(3.28)**	**(1.39)**	**(1.00)**
FP	[3.12]	[2.64]	[2.00]	[2.51]	[2.57]	[4.00]	[4.00]	[1.38]	[1.00]
	(3.24)	(2.04)	(3.56)	(2.04)	(2.72)	(2.00)	(2.00)	(1.33)	(1.00)

Table 3: Profits. Bold values correspond to the highest profits in the column with respect to each auction. Values within square brackets [·] refers to UA, while normal brackets (·) regards DA.

	EWA	*GW*	*QL*	*MM*	*avg*	*FP*	*2 pl.*	*3 pl.*	*4 pl.*
EWA	**[1.00]**	**[0.97]**	[0.71]	[0.72]	[0.85]	**[1.00]**	**[1.00]**	**[1.00]**	**[1.00]**
	(1.00)	(0.89)	(0.98)	**(0.97)**	**(0.96)**	**(1.00)**	**(1.00)**	**(1.00)**	**(1.00)**
GW	[0.97]	[0.91]	**[0.80]**	**[0.87]**	**[0.89]**	[0.80]	[0.91]	[0.97]	**[1.00]**
	(0.89)	**(0.94)**	(0.76)	(0.94)	(0.88)	(0.83)	(0.94)	**(1.00)**	**(1.00)**
QL	[0.71]	[0.80]	[0.56]	[0.62]	[0.67]	**[1.00]**	[0.56]	[0.78]	**[1.00]**
	(0.98)	(0.76)	**(1.00)**	(0.86)	(0.90)	**(1.00)**	**(1.00)**	**(1.00)**	(0.95)
MM	[0.72]	[0.87]	[0.62]	[0.59]	[0.70]	**[1.00]**	[0.59]	[0.57]	**[1.00]**
	(0.97)	**(0.94)**	(0.86)	(0.92)	(0.92)	(0.99)	(0.92)	(0.90)	**(1.00)**
FP	[1.00]	[0.80]	[1.00]	[1.00]	[0.83]	[0.99]	[0.99]	[0.74]	[1.00]
	(1.00)	(0.83)	(1.00)	(1.00)	(0.96)	(1.00)	(1.00)	(1.00)	(0.95)

Table 4: Frequencies of one-stage Nash equilibria in pure strategies. Bold values correspond to the highest frequency values in the column with respect to each auction. Values within square brackets [·] refers to UA, while normal brackets (·) regards DA.

	EWA	*GW*	*QL*	*MM*	*avg*	*FP*	*2 pl.*	*3 pl.*	*4 pl.*
EWA	**[1.00]**	**[1.00]**	**[0.99]**	[0.98]	**[0.99]**	**[1.00]**	**[1.00]**	**[0.69]**	**[0.00]**
	(0.64)	(0.16)	(0.61)	**(0.77)**	**(0.54)**	(0.62)	**(0.64)**	(0.00)	**(0.00)**
GW	**[1.00]**	**[1.00]**	[0.93]	**[1.00]**	[0.98]	**[1.00]**	**[1.00]**	[0.00]	**[0.00]**
	(0.16)	**(0.31)**	(0.01)	(0.07)	(0.14)	(0.02)	(0.31)	(0.00)	**(0.00)**
QL	[0.99]	[0.93]	[0.61]	[0.91]	[0.86]	**[1.00]**	[0.61]	[0.16]	**[0.00]**
	(0.60)	(0.01)	(0.45)	(0.67)	(0.43)	**(0.78)**	(0.45)	(0.00)	**(0.00)**
MM	[0.98]	**[1.00]**	[0.91]	[0.97]	[0.96]	[0.99]	[0.97]	[0.25]	**[0.00]**
	(0.77)	(0.07)	**(0.67)**	(0.63)	**(0.54)**	(0.02)	(0.63)	**(0.04)**	**(0.00)**
FP	[1.00]	[1.00]	[1.00]	[0.99]	[1.00]	[1.00]	[1.00]	[0.00]	[0.00]
	(0.62)	(0.02)	(0.78)	(0.02)	(0.36)	(0.00)	(0.00)	(0.00)	(0.00)

Table 5: Frequencies of one-stage strong Pareto optima. Bold values correspond to the highest frequency values in the column with respect to each auction. Values within square brackets [·] refers to UA, while normal brackets (·) regards DA.

	EWA	*GW*	*QL*	*MM*	*avg*	*FP*	*2 pl.*	*3 pl.*	*4 pl.*
EWA	**[0.00]**	[0.00]	[0.04]	[0.02]	[0.02]	**[0.00]**	[0.00]	**[0.00]**	**[0.00]**
	(0.00)	(0.26)	(0.04)	(0.01)	(0.08)	**(0.00)**	**(0.00)**	(0.08)	**(0.00)**
GW	**[0.00]**	**[-0.01]**	**[0.00]**	**[0.00]**	**[0.00]**	**[0.00]**	**[-0.01]**	**[0.00]**	**[0.00]**
	(0.00)	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**
QL	[0.23]	[0.38]	[0.78]	**[0.00]**	[0.35]	**[0.00]**	[0.78]	[0.11]	**[0.00]**
	(0.35)	(0.08)	(0.04)	(0.00)	(0.12)	**(0.00)**	(0.04)	(0.58)	**(0.00)**
MM	[0.38]	[0.02]	[0.34]	[0.20]	[0.24]	**[0.00]**	[0.20]	[0.17]	**[0.00]**
	(0.08)	(0.01)	(0.28)	(0.05)	(0.11)	**(0.00)**	(0.05)	(0.07)	**(0.00)**
FP	[0.00]	[0.00]	[0.00]	[0.00]	[0.12]	[0.00]	[0.00]	[0.23]	[0.00]
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00	(0.00)	(0.00)	(0.00)

Table 6: Average regrets. Bold values correspond to the smallest values in the column with respect to each auction. Values within square brackets [·] refers to UA, while normal brackets (·) regards DA.

	EWA	*GW*	*QL*	*MM*	*avg*	*FP*	*2 pl.*	*3 pl.*	*4 pl.*
EWA	**[0.00]**	**[0.01]**	[0.06]	[0.01]	[0.02]	**[0.00]**	**[0.00]**	**[0.00]**	**[0.00]**
	(0.00)	(0.04)	**(0.00)**	(0.01)	**(0.01)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**
GW	**[0.00]**	**[0.01]**	**[0.03]**	**[0.00]**	**[0.01]**	**[0.00]**	[0.01]	**[0.00]**	**[0.00]**
	(0.00)	**(0.01)**	(0.04)	**(0.00)**	**(0.01)**	**(0.00)**	(0.01)	**(0.00)**	**0.00**
QL	**[0.00]**	[0.02]	[0.10]	[0.01]	[0.03]	**[0.00]**	[0.10]	[0.18]	**[0.00]**
	(0.00)	(0.06)	**(0.00)**	(0.02)	(0.02)	**(0.00)**	**(0.00)**	(1.25)	(0.01)
MM	[0.07]	[0.04]	[0.08]	[0.08]	[0.07]	**[0.00]**	[0.08]	[0.22]	**[0.00]**
	(0.01)	(0.03)	(0.02)	(0.01)	(0.02)	**(0.00)**	(0.01)	(0.07)	**(0.00)**
FP	[0.00]	[0.04]	[0.00]	[0.00]	[0.03]	[0.00]	[0.00]	[0.33]	[0.00]
	(0.00)	(0.07)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)	(0.02)

Table 7: Average incentives to deviate. Bold values correspond to the smallest values in the column with respect to each auction. Values within square brackets [·] refers to UA, while normal brackets (·) regards DA.

Acknowledgments

This work has been partially supported by the University of Genoa, by the Italian Ministry of Education, University and Research (MIUR) under grants FIRB 2007.

A. Performance Metrics

The average regret metric measures the maximum average payoff loss of agent ith for playing the sequence of actions a_t^i instead of playing a fixed action s^i for every round t given that the opponents played the sequence a_t^{-i}. The average regret r_t^i is then defined as:

$$r_t^i = \max_{s^i} \frac{1}{t} \sum_{k=1}^{t} (R^i(s^i, a_k^{-i}) - R^i(a_k^i, a_k^{-i})) \tag{12}$$

Negative average regret means the ith agent's sequence of strategies outperformed every attainable fixed strategy s^i.
The average incentive to deviate (henceforth ID) d_t^i of an agent i gives the average payoff loss at round t for playing the actions a_t^i instead of playing the best response b_t^i given that the opponents played actions a_t^{-i}, i.e.,

$$d_t^i = \frac{1}{t} \sum_{k=1}^{t} R^i(b_k^i, a_k^{-i}) - R^i(a_k^i, a_k^{-i}) \tag{13}$$

It is worth noting that average regret and incentives to deviate measure different properties. Average regret measures the level of dissatisfaction of not having played a fixed action, i.e., a pure strategy, throughout all rounds by a learning agent. Whereas, the ID estimates the level of dissatisfaction of not having played a arbitrary non-stationary mixed strategy at every round.
A specific vector of strategies $a_* = (a_*^i, a_*^{-i})$ is a Nash equilibrium if the following conditions are satisfied:

$$R^i(a_*^i, a_*^{-i}) \geq R^i(a^i, a_*^{-i}), \text{ for any } i$$

In other terms, the previous formula states that a_* is a Nash equilibrium if no player has incentive to unilaterally change its action.
A Pareto optimum is a vector of actions $a_* = (a_*^i, a_*^{-i})$ for which there is not any other feasible vector of actions, say a, such that the solution a is strictly preferred by at least one player, and weakly preferred by everyone else. Formally, a specific vector of actions a_* is not a Pareto optimum if there exists another joint strategy a that satisfies the following conditions:

$$\begin{cases} R^i(a) \geq R^i(a_*), \text{ for any } i, \\ R^i(a) > R^i(a_*), \text{ at least for one } i \end{cases}$$

References

Bowling, M., 2005. Convergence and no-regret in multiagent learning. In: Advances in Neural Information Processing Systems. Vol. 17. MIT press, pp. 209–216.

Brown, G., 1951. Iterative solution of games by fictitious play. In: T.C. Koopmans, E. (Ed.), Activity Analysis of Production and Allocation. Wiley: New York, pp. 374–376.

Bunn, D. W., Oliveira, F. S., October 2001. Agent-based simulation - an application to the new electricity trading arrangements of england and wales. IEEE Trans Evolut Comput 5 (5), 493–503.

Camerer, C., 2003. Behavioral Game Theory. Princeton University Press.

Camerer, C., Ho, T., 1999. Experience-weighted attraction learning in normal-form games. Econometrica 67, 827–74.

Chang, Y., Kaelbling, L., 2001. Playing is believing: the role of beliefs in multi-agent learning. In: In Proceedings of NIPS-2001.

Cramton, P., October 1998. The efficiency of the FCC spectrum auctions. The Journal of law & economics 41 (2), 727–736.

Cramton, P., Kerr, S., March 2002. Tradeable carbon permit auctions - how and why to auction not grandfather. Energy policy 30 (4), 333–345.

Erev, I., Roth, A. K., September 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. Am Econ Rev 88 (4), 848–881.

Fabra, N., 2006. Designing electricity auctions. The Rand Journal of Economics 37 (1), 23–46.

Fudenberg, D., Levine, D., 2007. An economist's perspective on multi-agent learning. Artificial Intelligence 171 (7), 378–381.

Fudenberg, D., Levine, D. K., 1999. The Theory of Learning in Games. The MIT Press.

Gordon, G., 2007. Agendas for multi-agent learning. Journal of Artificial Intelligence 171 (7), 392–401.

Guerci, E., Ivaldi, S., Cincotti, S., in press. Learning agents in an artificial power exchangge: Tacit collusion, market power and efficiency of two double-auction. Computational Economics.

Guerci, E., Ivaldi, S., Raberto, M., Cincotti, S., May 2007. Learning oligopolistic competition in electricity auctions. Journal of Computational Intelligence 23 (2), 197–220.

Hu, J., Wellman, M., 1998. Multiagent reinforcement learning: theoretical framework and an algorithm. In: Proc. 15th International Conf. on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 242–250.

Klemperer, K., 2000. Auction theory: A guide to the literature. In: Cheltenham, U. (Ed.), The Double Auction Market: Institutions, Theories and Evidence. Edward Elgar, pp. 3–62.

Lipson, A., Leyton-Brown, K., 2005. Empirically evaluating multiagent reinforcement learning algorithms, working Paper.

Littman, M., 2001. Friend-or-Foe Q-learning in general-sum games. In: Proc. of 18th International Conference on Machine Learning. pp. 322–328.

Marimon, R., McGrattan, E., 1995. On adaptive learning in strategic games. In: Kirman, A., Salmon, M. (Eds.), Learning and Rationality in Economics. Blackwell, pp. 63–101.

Marks, E., 2006. Market design using agent-based models. In: Tesfatsion, L., Judd, K. (Eds.), Handbook of Computational Economics. Vol. 2. North Holland, pp. 1339–1380.

McKelvey, R., A.M.McLennan, Turocy, T., 2007. Gambit: Software tools for game theory. http://gambit.sourceforge.net, version 0.2007.01.30.

Milgrom, P., 1998. Game theory and the spectrum auctions. European Economic Review 42, 771–778.

Milgrom, P., 2004. Putting Auction Theory to Work. Cambridge University Press.

Nicolaisen, J., Petrov, V., Tesfatsion, L., October 2001. Market power and efficiency in a computational electricity market with discriminatory double-auction pricing. IEEE T Evolut Comput 5 (5), 504–523.

Roth, A. E., July 2002. The economist as engineer: gane theory, experimentation and computation as tools for design economics. Econometrica 70 (4), 1341–1378.

Sandholm, T., A. Gilpin, V., 2005. Mixed integer programming methods for finding nash equilibria. In: National Conference on Artificial Intelligence.

Shoham, Y., Powers, R., Grenager, T., 2007. If multi-agent learning is the answer, what is the question? Journal of Artificial Intelligence 171 (7), 365–377.

Singh, S., Kearns, M., Mansour, Y., 2000. Nash convergence of gradient dynamics in general-sum games. In: Proc. 16th Conference on Uncertainty in Artificial Intelligence. pp. 541–558.

Sun, J., Tesfatsion, L., 2007. Dynamic testing of wholesale power market designs: An open-source agent-based framework. Computational Economics 30 (3), 291–327.

Sutton, R. S., Barto, A. G., 1998. Reinforcement Learning: An introduction. The MIT press Cambridge.

Tesfatsion, L., Judd, K., 2006. Handbook of Computational Economics: Agent-Based Computational Economics. Vol. 2 of Handbook in Economics Series. North Holland.

Vohra, R., Wellman, M., 2007. Special issue on foundations of multi-agent learning. Artificial Intelligence 171 (7).

von der Fehr, N., Harbord, D., May 1993. Spot market competition in the UK electricity industry. Economic Journal 103, 531–546.

Watkins, C., Dayan, P., May 1992. Q-learning. Machine Learning 8 (3-4), 279–292.

Yu, N., Liu, C., Tesfatsion, L., 2007. Modeling of suppliers' learning behaviors in an electricity market environment. In: to appear in the Proceedings of the 14th International Conference on Intelligent System Applications to Power Systems (ISAP2007).